Multi-Stage Pre-training Enhanced by ChatGPT for Multi-Scenario Multi-Domain Dialogue Summarization

Weixiao Zhou†... Feifei Zhai... Zhoujun Li*†
†State Key Lab of Software Development Environment, Beihang University
‡State Key Lab of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS
§School of Artificial Intelligence, University of Chinese Academy of Sciences
¶Fanyu AI Research, Zhongke Fanyu Technology Co., Ltd
{wxzhou,buaacxf,xnliang,lizj}@buaa.edu.cn junnan.zhu@nlpr.ia.ac.cn

Abstract

Dialogue summarization involves a wide range of scenarios and domains. However, existing methods generally only apply to specific scenarios or domains. In this study, we propose a new pre-trained model specifically designed for multi-scenario multi-domain dialogue summarization. It adopts a multi-stage pre-training strategy to reduce the gap between the pre-training objective and fine-tuning objective. Specifically, we first conduct domain-aware pre-training using large-scale multi-scenario multi-domain dialogue data to enhance the adaptability of our pre-trained model. Then, we conduct task-oriented pre-training using large-scale multi-scenario multi-domain "dialogue-summary" parallel data annotated by ChatGPT to enhance the dialogue summarization ability of our pre-trained model. Experimental results on three dialogue summarization datasets from different scenarios and domains indicate that our pre-trained model significantly outperforms previous state-of-the-art models in full fine-tuning, zero-shot, and few-shot settings.1

1 Introduction

Dialogue summarization is the task of generating a summary from a dialogue (Xu et al., 2022). Specifically, open-domain dialogue summarization involves various scenarios (e.g., Online-Chat (Gliwa et al., 2019) and Daily-Life (Chen et al., 2021)), while customer service dialogue summarization involves different domains (e.g., Tweet (Feigenblat et al., 2021) and E-commerce (Lin et al., 2022)).

Recently, general-purpose pre-trained models have achieved significant success in dialogue summarization tasks (Lewis et al., 2020; Bao et al., 2020; Beltagy et al., 2020). Furthermore, several task-specific pre-trained models (Zhang et al., 2020a; Zhong et al., 2022) have further improved dialogue summarization. Existing dialogue summarization pre-trained model (Zhong et al., 2022) achieves good performance on long dialogue summarization. However, it still has the following limitations: (1) It is only pre-trained on dialogue corpora that include two domains (i.e., Interview and TV show), making it difficult to apply to dialogue summarization in a wide range of scenarios and domains. (2) It utilizes a window-based denoising task as the pre-training objective, which presents a significant gap with the fine-tuning objective. Simultaneously, existing state-of-the-art (SOTA) models generally improve dialogue summarization by modeling dialogue interactions (Lin et al., 2022; Tang et al., 2022), incorporating extra information (e.g., topics and roles) (Wang et al., 2022c; Kim et al., 2022), and rewriting dialogues (Xu et al., 2022; Fang et al., 2022). Although these methods have some effect, they still have limited applicability to downstream datasets in different scenarios and domains and are often difficult to apply within the current pre-training paradigm due to complex model architectures.

To address the limitations of previous works, in this study, our goal is to propose a task-specific pre-trained model for dialogue summarization, which has extremely small gap between the pre-training objective and the fine-tuning objective, enabling it to excellently adapt to downstream datasets from a wide range of scenarios and domains in full fine-tuning, few-shot, and zero-shot settings.1

1Our corpus and enhanced pre-trained models can be found at https://github.com/zhouweixiao/MP4
*Corresponding Author
We evaluate our pre-trained model on open-domain dialogue summarization datasets from two scenarios (i.e., Online-Chat (Gliwa et al., 2019) and Daily-Life (Chen et al., 2021)), as well as a customer service dialogue summarization dataset from a specific domain (i.e., Tweet (Feigenblat et al., 2021)). The experimental results indicate that MP4 significantly outperforms previous SOTA models in full fine-tuning, zero-shot, and few-shot settings, demonstrating remarkable performance improvements.

Our contributions are summarized as follows:

- We construct LCM³DS, which includes a large-scale collection of multi-scenario multi-domain dialogues and their corresponding summaries annotated by ChatGPT.
- We propose MP4, a multi-stage pre-trained model for multi-scenario multi-domain dialogue summarization.
- Our pre-trained model achieves new state-of-the-art performance on three dialogue summarization datasets from different scenarios and domains in full fine-tuning, zero-shot, and few-shot settings.

2 LCM³DS Corpus

2.1 Dialogue Preparation

Dialogue Collection. We collect 20 high-quality human-to-human multi-turn dialogue datasets to construct the dialogue part of LCM³DS, including 14 open-domain dialogue datasets across multiple scenarios (Rashkin et al., 2019; Nie et al., 2021; Dinan et al., 2019; Komeili et al., 2022; Gopalakrishnan et al., 2019; Zhang et al., 2018; Smith et al., 2020; Dinan et al., 2018; Li et al., 2017; Cui et al., 2020; Zhou et al., 2018; Yu et al.,
2020; Yang and Choi, 2019; Sun et al., 2019) and 6 multi-domain customer service dialogue datasets (Lee et al., 2019; Rastogi et al., 2020; Byrne et al., 2019; Zhang et al., 2020; Li et al., 2018; Feng et al., 2020). In total, there are 105,426 open-domain dialogues containing over 1M utterances and 101,342 customer service dialogues containing over 1.4M utterances. We provide the details of data statistics for all dialogue datasets in Appendix G.

### Dialogue Pre-Processing and Cleaning

We conduct a series of automated data pre-processing and cleaning to further improve the quality of the dialogues. For pre-processing, we perform the following steps: (1) Normalizing punctuations, special characters, and capitalization in each dialogue. (2) Following previous studies (Dinan et al., 2019; Chen et al., 2021), we preprocess each dialogue into a dual-turn dialogue format by merging consecutive utterances from the same speaker. For cleaning, we perform the following steps: (1) Removing duplicate and highly similar dialogues using the Jaccard text similarity algorithm. (2) Deleting highly similar dialogues between the dialogue datasets and the evaluation datasets using the same algorithm as in (1), ensuring that they have no intersection. (3) Removing dialogues with less than 4 turns or 32 tokens.

### Role Adding

In order to standardize the different speaker formats of original dialogues across various datasets, we collect a list containing over 4,000 real names. For each dialogue, we randomly selected several real names from the list to assign a role group (e.g., Danny and Alejandra), where the number and order of real names in each role group corresponds to the speakers in the original dialogue.

### 2.2 Annotation

**Prompt Format.** We follow the previous study of InstructGPT (Ouyang et al., 2022) by inserting the text "Tldr: " at the end of each dialogue as a prompt and inputting it into ChatGPT (in zero-shot setting) to obtain annotated summaries. We also investigate the performance of three different prompts for dialogue summarization in zero-shot setting, and the details can be found in Appendix A.

### Role-Replaced Data Augmentation

In dialogue summarization, there are multiple scenarios and domains involving different roles. To alleviate this problem, we propose a simple yet effective method that can be extended to dialogue summarization involving any role. Specifically, we directly replace the roles in the dialogues and summaries from LCM3DS to obtain an augmented parallel corpus. In this study, we perform replacements for two common types of roles, including named coreference and customer service. The example we provide can be found in Appendix C.

### 2.3 Data Analysis

We empirically compare LCM3DS with existing dialogue summarization datasets (Gliwa et al., 2019; Chen et al., 2021; Feigenblat et al., 2021) based on five metrics (Grusky et al., 2018; Fabbri et al., 2021): Compression Ratio, Coverage, Density, Novelty, and Redundancy (see Table 1).

Compared to existing dialogue summarization datasets, LCM3DS exhibits lower Compression Ratio, moderate Coverage, and lower Density, indicating that the summaries maintain a high degree of abstraction while covering the important content of the dialogue and retaining more details and information from the dialogue. Additionally, LCM3DS shows higher Novelty and Redundancy, which is mainly caused by the lower Compression Ratio. Furthermore, unlike existing small-scale, human-annotated, single-scenario, single-domain dialogue summarization datasets, LCM3DS is large-scale, ChatGPT-annotated, multi-scenario, and multi-domain.

---

3We use gpt-3.5-turbo-0301
3 Model

3.1 Dialogue Modeling

MP4 is based on the standard Transformer (Vaswani et al., 2017) encoder-decoder architecture. For multi-turn dialogue modeling, the input embedding of each token is the sum of the corresponding token, position, and speaker embeddings. Figure 2 illustrates the dialogue modeling of MP4.

**Input Structure.** Given the dialogue context $D$, we first concatenate all roles $R_i$ and utterances $U_j$ in the dialogue context with two additional special tokens as a separate, consecutive token sequence: $X = \{R_1<\text{eor}>U_1<\text{eou}>\ldots R_m<\text{eor}>U_n<\text{eou}>\}$ where the special end-of-role token $<\text{eor}>$ and end-of-utterance token $<\text{eou}>$ are respectively appended to the end of each role and utterance for multi-turn dialogue separation. Then, we add the start token $<s>$ and the end token $</s>$ around the token sequence $X$ as the input for MP4.

**Speaker Embeddings.** To distinguish utterances in the dialogue context and capture underlying role interactions during the dialogue process, we follow previous dialogue modeling studies (Gu et al., 2020, 2021) and add additional speaker embeddings to token representations. This process is performed alternately based on the role transitions and can be extended to dialogues with an unlimited number of roles. The speaker embeddings are combined with the initial token and position embeddings and then fed into the MP4 encoder-decoder framework.

3.2 Multi-Stage Pre-training

We conduct multi-stage pre-training to reduce the gap between the pre-training objective and the fine-tuning objective. Domain-aware pre-training aims to enhance the adaptability of MP4 to dialogues in multiple scenarios and domains, while task-oriented pre-training aims to enhance the ability of MP4 to summarize unstructured spoken multi-scenario multi-domain dialogues into structured written-language summaries.

3.2.1 Domain-Aware Pre-training

General-purpose pre-trained models (Lewis et al., 2020) are pre-trained on free-form text data with universal pre-training objectives, limiting their ability in specific domains and tasks. Therefore, it is common practice to further train these models with the language modeling objective using text from the target domain to reduce negative impact (Zhang and Zhao, 2021; Whang et al., 2021). In this study, we conduct a domain-aware pre-training stage on MP4 using the dialogue data from LCM$^3$Ds. Specifically, we achieve this by modeling a series of dialogue reconstruction pre-training objectives inspired by BART (Lewis et al., 2020).

**Token Masking.** For tokens of each utterance in the dialogue, 20% of them are randomly sampled and replaced with a special $<\text{mask}>$ token.

**Token Deletion.** 20% of the tokens in the dialogue utterances are randomly sampled and deleted.

**Utterance Infilling.** Several utterance spans are randomly sampled, and each span is replaced with a single $<\text{mask}>$ token. The length of each utterance span is drawn from the Poisson Distribution ($\lambda = 3$). 0-length spans correspond to the insertion of $<\text{mask}>$ tokens.

**Utterance Permutation.** The order of all utterances in the dialogue turns is randomly shuffled. In contrast to previous studies (Zhong et al., 2022; Wang et al., 2022b), we did not shuffle the order of roles. Therefore, MP4 needs to reconstruct the correct order of utterances and ensure the precise alignment between utterances and roles.

**Utterance Masking.** 20% of the utterances in the dialogue are selected and replaced with a special $<\text{uttr-mask}>$ token. We did not perform random
selection but instead followed the method of PEGASUS (Zhang et al., 2020a) using greedy search to obtain the principal Gap-utterances. During the decoding process, MP4 needs to reconstruct the complete dialogue.

**Multi-Task Learning.** The model is trained with a maximum likelihood objective $L_{\Theta}$. Given the training sample $D = (x, y)$, $L_{\Theta}$ is defined as

$$L_{\Theta} = -\sum_{i=1}^{|y|} \log P_{\Theta}(y_i|y_{<i}; x)$$  \hspace{0.5cm} (1)

where $\Theta$ is the model parameters, $x$ is the noisy dialogue, and $y$ is the original dialogue.

During each iteration of the multi-task domain-aware pre-training stage, training samples are randomly selected from different pre-training tasks as mini-batches and used to calculate the cumulative loss and optimize the model parameters $\Theta$.

### 3.2.2 Task-Oriented Pre-training

Several task-specific summarization pre-trained models (Zhang et al., 2020a; Xiao et al., 2022; Zhong et al., 2022) reduce the gap with downstream datasets by modeling the task-oriented pre-training objective. Specifically, they typically select segments (e.g., gap-sentences or window-based utterances) of the original text (e.g., document or dialogue) as optimization targets for the decoder. Although they have some effects, however, there still exists a significant gap between the segments selected through unsupervised methods and abstractive written-language summaries. In this study, we directly utilize the "dialogue-summary" parallel data from LCM4Ds for task-oriented pre-training stage. The learning objective is similar to Eq. (1), where the training sample $D = (x, y)$, with $x$ representing the original dialogue and $y$ representing the summary annotated by ChatGPT.

### 4 Experiments

#### 4.1 Experimental Setup

**Implementation Details.** MP4 is initialized with BART-large\(^3\) (Lewis et al., 2020), which is a denoising sequence-to-sequence pre-trained Transformer (Vaswani et al., 2017) model with 12 layers and 16 attention heads. To facilitate performance comparison, we have implemented four types of MP4 models. MP4 (VANILLA) represents the non-pretrained model, which includes only specialized input structure and speaker embeddings. MP4 (DAP) denotes domain-aware pre-training applied to MP4 (VANILLA). MP4 (TOP) signifies task-oriented pre-training applied to MP4 (VANILLA). MP4 (DAP-TOP) indicates multi-stage pre-training applied to MP4 (VANILLA). More implementation details of pre-training are provided in Appendix D.

**Downstream Datasets.** We evaluate the performance of MP4 on open-domain dialogue summarization datasets from two scenarios (i.e., Online-Chat and Daily-Life), namely SAMSum (Gliwa et al., 2019) and DIALOGSUM (Chen et al., 2021), as well as a customer service dialogue summarization dataset from a specific domain (i.e., Tweet), namely TWEETSUMM (Feigenblat et al., 2021). Table 1 provides the statistics of the downstream datasets. More details are provided in Appendix B.

**Comparison Methods.** We compare MP4 with three types of baselines: extractive models, abstractive models, and previous SOTA models. The following presents the comparison methods.

- **Extractive Models.** Based on heuristic algorithms or graph-based algorithms, including LONGEST, Lead-3, and TextRank (Mihalcea and Tarau, 2004).

- **Abstractive Models.** Based on neural network sequence-to-sequence models, including PGNet (See et al., 2017), FastAbs-RL (Chen and Bansal, 2018), and Transformer (Vaswani et al., 2017).

- **Previous SOTA Models.** State-of-the-art dialogue summarization models based on pre-trained models, including BART(\(D_{\text{ALL}}\)) (Feng et al., 2021), Coref-ATTN (Liu et al., 2021), BART-ConFiT (Tang et al., 2022), DialSent-PGG (Jia et al., 2022), BART-NARR (Xu et al., 2022), MV-BART (Chen and Yang, 2020), ReWriteSum (Fang et al., 2022), BART-SCL (Geng et al., 2022), UniLMv2 (Bao et al., 2020), BART (Lewis et al., 2020), LA-BART (Wang et al., 2022a), and BART-MT (Bhattacharjee et al., 2022).

**Evaluation Metrics.** We evaluate the full fine-tuning, zero-shot, and few-shot performance of all models using ROUGE scores\(^4\) (i.e., R-1, -2, and -L), which are standard evaluation metrics.

---

\(^3\)https://huggingface.co/facebook/bart-large

\(^4\)https://pypi.org/project/py-rouge
### 4.2 Full Fine-Tuning Evaluation

To demonstrate the advantages of our pre-trained model with a large amount of training samples, we train the model using the entire training set for full fine-tuning evaluation.

**Settings.** We provide all the hyper-parameters used for fine-tuning and inference in Appendix E. During the evaluation, for SAMSUM, we followed (Liu and Lapata, 2019) by testing with the top-3 best checkpoints on the validation set and reporting the average ROUGE scores. For DIALOGSUM, we followed (Chen et al., 2021) by reporting the average ROUGE scores between the inference output and multiple reference summaries. For TWEETSUMM, due to limited research and the lack of detailed evaluation procedures in the original paper (Feigenblat et al., 2021), we suggest the following evaluation method: (1) During training, use the reference summary with the highest ROUGE-Avg score (i.e., the average value of R-1, R-2, and R-L) between the original dialogue and multiple reference summaries for training. (2) During testing, calculate the average ROUGE scores between the inference output and multiple reference summaries.

**Results.** Since most previous SOTA models have not been evaluated on a wide range of dialogue summarization datasets, therefore, we present the full fine-tuning performance on the SAMSUM, DIALOGSUM, and TWEETSUMM test sets in Tables 2, 3, and 4, respectively. Compared to previous SOTA models, our MP4 (DAP-TOP) achieves new state-of-the-art results on all metrics across three downstream datasets from different scenarios and domains, demonstrating significant performance improvements. Specifically, on SAMSUM and DIALOGSUM, MP4 (DAP-TOP) surpasses the previous SOTA models by improving the R-2 score by 0.70 and 0.54 (29.87→30.57 and 21.18→21.72), respectively. The improvement on TWEETSUMM reaches 1.46 (22.14→23.60). This indicates that multi-stage pre-training can assist the model better adapt to downstream datasets from a wide range of scenarios and domains. Additionally, compared to BART-large, MP4 (VANILLA) demonstrates stronger performance on most metrics, proving the effectiveness of introducing dialogue modeling. Furthermore, compared to MP4 (VANILLA), the results of MP4 (DAP) indicate that domain-aware pre-training can enhance the adaptability of model to dialogues in multiple scenarios and domains. Moreover, MP4 (TOP) achieves significant performance improvements, highlighting the importance of equipping the model with the ability to summarize multi-scenario multi-domain dialogues during the pre-training stage.

---

**Table 2:** Full fine-tuning results on SAMSUM test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous SOTA Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART-ATTN (Feng et al., 2021)</td>
<td>53.70</td>
<td>28.79</td>
<td>50.81</td>
</tr>
<tr>
<td>Coref-ATTN (Liu et al., 2021)</td>
<td>53.93</td>
<td>28.58</td>
<td>50.39</td>
</tr>
<tr>
<td>BART-ConfIT (Tang et al., 2022)</td>
<td>53.89</td>
<td>28.85</td>
<td>49.29</td>
</tr>
<tr>
<td>DialSent-PGG (Jia et al., 2022)</td>
<td>53.54</td>
<td>28.91</td>
<td>50.21</td>
</tr>
<tr>
<td>BART-NARR (Xu et al., 2022)</td>
<td>53.80</td>
<td>28.96</td>
<td>50.76</td>
</tr>
<tr>
<td>MV-BART (Feng et al., 2022)</td>
<td>54.05</td>
<td>28.56</td>
<td>50.57</td>
</tr>
<tr>
<td>ReWriteSum (Fang et al., 2022)</td>
<td>54.20</td>
<td>27.10</td>
<td>50.10</td>
</tr>
<tr>
<td>BART-SCL (Geng et al., 2022)</td>
<td>54.22</td>
<td>29.87</td>
<td>51.35</td>
</tr>
<tr>
<td>Our Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART-large</td>
<td>53.32</td>
<td>28.78</td>
<td>50.63</td>
</tr>
<tr>
<td>MP4 (VANILLA)</td>
<td>53.41</td>
<td>29.23</td>
<td>50.97</td>
</tr>
<tr>
<td>MP4 (DAP)</td>
<td>53.82</td>
<td>29.55</td>
<td>51.21</td>
</tr>
<tr>
<td>MP4 (TOP)</td>
<td>54.56</td>
<td>30.33</td>
<td>51.70</td>
</tr>
<tr>
<td>MP4 (DAP-TOP)</td>
<td>54.60</td>
<td>30.57</td>
<td>51.79</td>
</tr>
</tbody>
</table>

**Table 3:** Full fine-tuning results on DIALOGSUM test set. We report the average of multiple-reference results.

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous SOTA Models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART-large</td>
<td>46.68</td>
<td>20.96</td>
<td>44.68</td>
</tr>
<tr>
<td>MP4 (VANILLA)</td>
<td>46.87</td>
<td>21.21</td>
<td>44.87</td>
</tr>
<tr>
<td>MP4 (DAP)</td>
<td>47.01</td>
<td>21.37</td>
<td>44.94</td>
</tr>
<tr>
<td>MP4 (TOP)</td>
<td>47.74</td>
<td>21.84</td>
<td>45.77</td>
</tr>
<tr>
<td>MP4 (DAP-TOP)</td>
<td>48.01</td>
<td>21.72</td>
<td>45.92</td>
</tr>
</tbody>
</table>

**Table 4:** Full fine-tuning results on TWEETSUMM test set. * denotes results obtained from (Feigenblat et al., 2021).
### 4.3 Zero- and Few-Shot Evaluation

Many existing studies that apply pre-trained models to dialogue summarization require a large amount of fine-tuning data, which is often impractical in new scenarios or domains. In contrast, we expect our model to quickly adapt to new scenarios or domains without the need for a large amount of fine-tuning data. To validate this hypothesis, we conduct evaluations in zero-shot (no training samples) and few-shot (10 training samples) settings. Obtaining such a small number of samples is feasible in practice for new scenarios or domains.

#### Settings.

We compare the performance of BART-large (Lewis et al., 2020), MP4 (DAP), and MP4 (DAP-TOP) in zero-shot and few-shot settings. In Appendix F, we provide all the hyper-parameters used. Specifically, for zero-shot evaluation, since the models have not been trained on downstream datasets, we report the results of using the optimal summary length limits during inference. For few-shot evaluation, we randomly sample 10 training samples for training. Additionally, to ensure that the results are not affected by sampling variability, we conduct the same experiment five times with different random seeds (shared among all models) and report the average results.

#### Results.

The results presented in Table 5 indicate that our pre-trained model achieves significant improvements compared to BART-large. Specifically, for zero-shot results, MP4 (DAP-TOP) increases the R-1 score by 14.47 (27.94 → 42.41), 13.26 (25.50 → 38.76), and 9.50 (29.42 → 38.92) on SAMSUm, DIALOGSUM, and TWEETSUMM, respectively. Moreover, the zero-shot performance of MP4 (DAP-TOP) surpassed the few-shot performance of BART-large on multiple datasets, demonstrating its powerful zero-shot capability. Additionally, the few-shot results also highlight the advantages of MP4 (DAP-TOP), indicating that our pre-trained model converges faster than other models even with only 10 training samples.

<table>
<thead>
<tr>
<th>Model</th>
<th>SAMSUm</th>
<th>DIALOGSUM</th>
<th>TWEETSUMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>zero-shot</td>
<td>few-shot (10)</td>
<td>zero-shot</td>
</tr>
<tr>
<td>BART-large (ours)</td>
<td>27.94/8.56/26.03</td>
<td>39.47/16.51/38.76</td>
<td>25.50/6.10/25.32</td>
</tr>
<tr>
<td>MP4 (DAP)</td>
<td>32.48/9.80/31.03</td>
<td>41.75/17.34/40.29</td>
<td>27.80/6.61/27.42</td>
</tr>
<tr>
<td>MP4 (DAP-TOP)</td>
<td>42.41/17.15/39.69</td>
<td>48.73/23.63/45.96</td>
<td>38.76/14.59/37.18</td>
</tr>
</tbody>
</table>

Table 5: R-1/R-2/R-L results in zero-shot and few-shot settings. For zero-shot setting, we report the results at the optimal summary length limits. For few-shot setting, we report the average results from 5 random runs on 10 training samples (all models share the same seed set).

### 4.4 Ablation Study

To further validate the contributions of the fine-grained components in our pre-trained models, we conduct an ablation study on SAMSUm in full fine-tuning setting. The token-level tasks refer to Token Masking and Token Deletion.

#### Model R-1 R-2 R-L

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP4 (VANILLA)</td>
<td>53.41</td>
<td>29.23</td>
<td>50.97</td>
</tr>
<tr>
<td>w/o speaker embeddings</td>
<td>53.29</td>
<td>29.02</td>
<td>50.85</td>
</tr>
<tr>
<td>MP4 (DAP)</td>
<td>53.82</td>
<td>29.55</td>
<td>51.21</td>
</tr>
<tr>
<td>w/o token-level tasks</td>
<td>53.74</td>
<td>29.25</td>
<td>51.05</td>
</tr>
<tr>
<td>w/o utterance infilling</td>
<td>53.75</td>
<td>29.31</td>
<td>51.06</td>
</tr>
<tr>
<td>w/o utterance permutation</td>
<td>53.69</td>
<td>29.38</td>
<td>51.08</td>
</tr>
<tr>
<td>w/o utterance masking</td>
<td>53.53</td>
<td>29.30</td>
<td>50.94</td>
</tr>
<tr>
<td>MP4 (TOP)</td>
<td>54.56</td>
<td>30.33</td>
<td>51.70</td>
</tr>
<tr>
<td>w/o CSDS pre-training corpus</td>
<td>54.47</td>
<td>30.12</td>
<td>51.69</td>
</tr>
<tr>
<td>MP4 (DAP-TOP)</td>
<td>54.60</td>
<td>30.57</td>
<td>51.79</td>
</tr>
</tbody>
</table>

Table 6: Ablation study on SAMSUm in full fine-tuning setting. The token-level tasks refer to Token Masking and Token Deletion.

#### Speaker Embeddings.

As the results show, incorporating additional speaker embeddings in dialogue modeling can capture the underlying role interactions during the dialogue process and improve the performance of dialogue summarization.

#### Domain-Aware Pre-training Objectives.

As the results show, each domain-aware pre-training objective brings performance improvements. It is worth noting that the utterance masking task has the greatest impact on performance, indicating that completing principal Gap-utterances during dialogue reconstruction is crucial for dialogue summarization.

#### Impact of CSDS Pre-training Corpus on ODDS.

We remove the customer service "dialogue-summary" parallel data from LCM3Ds to investigate the impact of this portion of data on the performance of open-domain dialogue summarization. The results show that the model trained without this data exhibits a slight decrease in performance. One possible reason is that a few dialogues in SAMSUm also involve customer service topics.
5 Human Evaluation

We conduct human evaluation to further evaluate the performance of our pre-trained model and strong baselines under various paradigms, as well as the Ground Truth (i.e., MP4 (DAP-TOP), ChatGPT, BART-large, Ground Truth). Specifically, we randomly select 50 samples from the test set of SAMSUm. Then, we invite 3 participants to rank four candidate summaries according to four metrics: fluency (Flu.), conciseness (Conci.), informativeness (Info.), and comprehensiveness (Comp.). The top-ranking indicates the best performance on that metric.

Table 7 shows the results of human evaluation (lower average rank is better). Our pre-trained model outperforms BART-large in all metrics but falls behind the Ground Truth. Specifically, ChatGPT achieves the first rank in fluency and comprehensiveness for the summaries generated in zero-shot setting, surpassing the Ground Truth. However, it exhibits the weakest performance in conciseness and informativeness. The main reason for this is that ChatGPT tends to generate longer summaries that describe various aspects of the dialogue, including both important and minor details. Moreover, the longer summaries also contribute to an improved overall impression to some extent.

6 Related Work

Dialogue Corpora. In general, dialogue data can be obtained from two main sources. One is massive-scale dialogue corpora crawled from web platforms such as Reddit (Baumgartner et al., 2020; Henderson et al., 2019) and Twitter (Ritter et al., 2010), which are commonly used for pre-training open-domain chatbots (Zhang et al., 2020b; Adiwardana et al., 2020; Roller et al., 2021; Chen et al., 2022; Henderson et al., 2020). Another source is a collection of open-source dialogue datasets designed for specific tasks, including open-domain dialogue systems (Li et al., 2017; Zhou et al., 2018; Dinan et al., 2018; Smith et al., 2020; Zhang et al., 2018; Gopalakrishnan et al., 2019; Komeili et al., 2022; Dinan et al., 2019), task-oriented dialogue systems (Lee et al., 2019; Rastogi et al., 2020; Byrne et al., 2019; Zhang et al., 2020; Li et al., 2018; Feng et al., 2020), dialogue comprehension (Sun et al., 2019; Yang and Choi, 2019; Yu et al., 2020; Cui et al., 2020; Nie et al., 2021; Rashkin et al., 2019), and dialogue summarization (Gliwa et al., 2019; Chen et al., 2021; Feigenblat et al., 2021).

PTMs for Dialogue Summarization. Recently, general-purpose pre-trained models have achieved significant success in dialogue summarization tasks (Lewis et al., 2020; Raffel et al., 2020; Bao et al., 2020; Beltagy et al., 2020). Furthermore, several task-specific pre-trained models (Zhang et al., 2020a; Zhong et al., 2022) have further improved dialogue summarization. Moreover, existing state-of-the-art dialogue summarization models typically leverage pre-trained models and model the characteristics of dialogues to achieve better results, including modeling dialogue interactions (Lin et al., 2022; Tang et al., 2022), incorporating extra information (Wang et al., 2022c; Kim et al., 2022), and dialogue rewriting (Xu et al., 2022; Fang et al., 2022). Although these models are effective, they often have complex model structures that are difficult to apply within the current pre-training paradigm.

Large Language Models. More recently, LLMs have attracted widespread attention due to their remarkable performance in various knowledge-intensive NLP tasks (Zeng et al., 2022; Ouyang et al., 2022; Thoppilan et al., 2022; Scao et al., 2022). Through large-scale pre-training on massive text corpora (Brown et al., 2020; Wei et al., 2022b), LLMs possess powerful foundational capabilities. Instruction tuning (Raffel et al., 2020; Wei et al., 2022a; Chung et al., 2022) helps LLMs in understanding natural language task descriptions. while Reinforcement Learning with Human Feedback (RLHF) (Stiennon et al., 2020; Bai et al., 2022) aligns generated text with human preferences.

7 Conclusion

In this study, we propose MP4, a multi-stage pre-trained model for multi-scenario multi-domain dialogue summarization. To conduct the pre-training, we construct a large-scale ChatGPT-annotated multi-scenario multi-domain multi-turn dialogue summarization corpus called LCM$^3$DS. Extensive experimental results demonstrate that MP4 exhibits remarkable dialogue summarization capabilities.
Limitations

Although we have demonstrated the powerful performance of MP4 in multi-scenario multi-domain dialogue summarization, there are still some limitations that provide directions for future work: (1) Due to limitations in computational resources, we did not consider long dialogues when constructing $\text{LCm}^3$Ds. Therefore, MP4 may be more suitable for short dialogue summarization. (2) MP4 is initialized with BART-large, which has only 0.4 billion parameters. In future work, we will consider using larger base models.

Ethics Statement

All dialogue data used in this study are sourced from previously published works, and there are no copyright restrictions on their academic use, allowing free online access. Since we used ChatGPT for data annotation, the $\text{LCm}^3$Ds corpus we constructed is intended solely for academic research purposes. Additionally, MP4 is initialized with the weights from BART-large. Therefore, MP4 may exhibit biases and harmful behaviors commonly observed in language models.

Acknowledgement

This work received support from both the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081) and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18). We extend our sincere appreciation to Yu Zhou and Rongping Chang for their valuable contributions to the revision of this paper.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot.


Xiaochong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers.


Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots.


Bin Wang, Chen Zhang, Chengwei Wei, and Haizhou Li. 2022a. A focused study on sequence length for dialogue summarization.


Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language
models are zero-shot learners. In International Conference on Learning Representations.


A Prompts for Dialogue Summarization

We utilize ChatGPT as an unsupervised summarizer to annotate dialogues in the SAMSum (Gliwa et al., 2019) test set using three different prompt formats in zero-shot setting:

- **Preceding Prompt.** Insert the prompt "Summarize the following dialogue into a short summary:" before the dialogue.

- **InstructGPT Prompt.** Following InstructGPT (Ouyang et al., 2022) insert the prompt "Tldr:" after the dialogue.

- **Subsequent Prompt.** Insert the prompt "Summarize the above dialogue into a short summary:" after the dialogue.

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preceding</td>
<td>37.90</td>
<td>15.19</td>
<td>35.89</td>
</tr>
<tr>
<td>InstructGPT</td>
<td>42.17</td>
<td>16.84</td>
<td>39.26</td>
</tr>
<tr>
<td>Subsequent</td>
<td>40.08</td>
<td>15.41</td>
<td>37.22</td>
</tr>
</tbody>
</table>

Table 8: Comparison of dialogue summarization performance for different prompts on SAMSum test set, which contains 819 dialogues.

Table 8 shows the evaluation results of three different prompts. The **InstructGPT Prompt** achieves the best performance.

B Details of Downstream Datasets

**SAMSum.** A natural messenger dialogue summarization dataset containing dialogues created and written down by linguists fluent in English. The dataset includes various topics such as chit-chats, gossiping about friends, arranging meetings, discussing politics, consulting university assignments with colleagues, etc. The sizes of the training, validation, and test sets are 14,731, 818, and 819 respectively.

**DIALOGSUM.** A large-scale dataset for dialogue summarization, which includes face-to-face spoken dialogues covering a wide range of daily-life topics, including schooling, work, medication, shopping, leisure, and travel, etc. Most dialogues take place between friends, colleagues, and between service providers and customers. The sizes of the training, validation, and test sets are 12,460, 500, and 500 respectively.

**TWEETSUMM.** A customer service dialogue summarization dataset, consisting of reconstructed dialogues extracted from the Kaggle Customer Support On Twitter dataset. The Kaggle dataset is a large-scale dataset based on dialogues between consumers and customer support agents on Twitter.com. It covers a wide range of topics and services provided by various companies, including airlines, retail, gaming, music, etc. The sizes of the training, validation, and test sets are 869, 108, and 110 respectively.

C Example of Data Augmentation

<table>
<thead>
<tr>
<th>Dialogue</th>
<th>Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Danny</td>
<td><img src="image" alt="Prompt" /></td>
</tr>
<tr>
<td>Are there any pg-13 rated movies playing around 6:00pm tomorrow at zip code 70072?</td>
<td>Are there any pg-13 rated movies playing around 6:00pm tomorrow at zip code 70072?</td>
</tr>
<tr>
<td>Alejandra</td>
<td><img src="image" alt="Prompt" /></td>
</tr>
<tr>
<td>Sure! There's Batman vs Superman at 6pm. The rest are later in the evening.</td>
<td>Are you interested in that?</td>
</tr>
<tr>
<td>Danny</td>
<td><img src="image" alt="Prompt" /></td>
</tr>
<tr>
<td>How many tickets?</td>
<td>Yes, that's fine at 6pm.</td>
</tr>
<tr>
<td>Alejandra</td>
<td><img src="image" alt="Prompt" /></td>
</tr>
<tr>
<td>code 70072?</td>
<td>How many tickets?</td>
</tr>
<tr>
<td>Danny</td>
<td><img src="image" alt="Prompt" /></td>
</tr>
<tr>
<td>4 tickets for the 6pm showing.</td>
<td>4 tickets for the 6pm showing.</td>
</tr>
</tbody>
</table>

Figure 3: An example of role-replaced dialogue and summary. Red indicates a role group added through role adding, yellow indicates named coreference, and blue indicates roles in customer service.

D Implementation Details of Pre-training

The following are the hyper-parameters used in domain-aware pre-training.

- **gpus** 8
- **steps** 5000
- **batch_size** 16
- **lr** 3e − 05
- **warmup_steps** 500
- **label_smoothing** 0.1
- **optimizer** Adam

The following are the hyper-parameters used in task-oriented pre-training.
E Details of Fine-Tuning and Inference

The following are the hyper-parameters used in full fine-tuning setting on SAMSum.

- `gpus` 4
- `steps` 1150
- `batch_size` 16
- `lr` 3e-05
- `warmup_steps` 100
- `label_smoothing` 0.1
- `optimizer Adam`

The following are the hyper-parameters used in full fine-tuning setting on DIALOGSUM.

- `gpus` 4
- `steps` 1000
- `batch_size` 16
- `lr` 3e-05
- `warmup_steps` 100
- `label_smoothing` 0.1
- `optimizer Adam`

The following are the hyper-parameters used in full fine-tuning setting on TWEETSUMM.

- `gpus` 4
- `steps` 98
- `batch_size` 16
- `lr` 3e-05
- `warmup_steps` 10
- `label_smoothing` 0.1
- `optimizer Adam`

All models maintain consistent hyper-parameters across all datasets during inference.

- `gpus` 1
- `batch_size` 32
- `use_cache` true
- `max_length` 100
- `min_length` 5
- `beam_size` 5
- `length_penalty` 1
- `no_repeat_ngram_size` 0
- `early_stopping` false

F Zero- and Few-Shot Evaluation Details

For zero-shot evaluation, the optimal summary length limit hyper-parameter `max_length` for SAMSum, DIALOGSUM, and TWEETSUMM is 60, 40, and 80 respectively. Moreover, other hyper-parameters used during inference remain consistent with Appendix E. For few-shot evaluation (with 10 training samples), we provide the hyper-parameter settings used during training below, while the hyper-parameters used during inference are consistent with Appendix E.

- `gpus` 1
- `steps` 20
- `batch_size` 10
- `lr` 3e-05
- `warmup_steps` 0
- `label_smoothing` 0.1
- `optimizer Adam`
- `seeds` 3442|3443|3444|3445|3446

G Data Statistics for Dialogue Datasets

Please refer to Table 9.

H Examples of Generated Summaries

Please refer to Figure 4.
Table 9: Data statistics for preprocessed, cleaned and annotated dialogue datasets of LCM\textsuperscript{4}Ds. # denotes average value. Dial. represents dialogue. Summ. represents ChatGPT-annotated summary. Sce. represents scenario of open-domain dialogue. Dom. represents domains of customer service dialogue.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Dialogue</th>
<th>#Utterance</th>
<th>Tokens/dial</th>
<th>Tokens/turn</th>
<th>Tokens/summ</th>
<th>Sce./Dom.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open-Domain Dialogue</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathetic Dialogues</td>
<td>22,878</td>
<td>98,759</td>
<td>4.3</td>
<td>86.5</td>
<td>20.1</td>
<td>34.2</td>
</tr>
<tr>
<td>Wizard of Wikipedia</td>
<td>19,287</td>
<td>174,624</td>
<td>9.1</td>
<td>213.7</td>
<td>23.5</td>
<td>55.8</td>
</tr>
<tr>
<td>Wizard of Internet</td>
<td>9,025</td>
<td>91,042</td>
<td>10.1</td>
<td>225.0</td>
<td>22.3</td>
<td>55.6</td>
</tr>
<tr>
<td>Topical-Chat (Gopalanakrishnan et al., 2019)</td>
<td>6,324</td>
<td>140,544</td>
<td>21.5</td>
<td>507.7</td>
<td>23.6</td>
<td>61.2</td>
</tr>
<tr>
<td>Persona-Chat (Zhang et al., 2018)</td>
<td>5,648</td>
<td>84,366</td>
<td>14.9</td>
<td>239.9</td>
<td>16.1</td>
<td>55.0</td>
</tr>
<tr>
<td>Blended Skill Talk (Smith et al., 2020)</td>
<td>4,662</td>
<td>62,581</td>
<td>13.4</td>
<td>268.1</td>
<td>20.0</td>
<td>53.8</td>
</tr>
<tr>
<td>ConvAI2 (Dinan et al., 2018)</td>
<td>4,554</td>
<td>67,532</td>
<td>14.8</td>
<td>237.8</td>
<td>16.1</td>
<td>53.9</td>
</tr>
<tr>
<td>MeTaTal (Cui et al., 2020)</td>
<td>2,719</td>
<td>14,134</td>
<td>5.2</td>
<td>98.7</td>
<td>19.0</td>
<td>38.1</td>
</tr>
<tr>
<td>CMU DoG (Zhou et al., 2018)</td>
<td>2,647</td>
<td>58,406</td>
<td>22.1</td>
<td>422.4</td>
<td>19.1</td>
<td>67.9</td>
</tr>
<tr>
<td>DailyDialog (Li et al., 2017)</td>
<td>2,471</td>
<td>13,891</td>
<td>5.6</td>
<td>79.0</td>
<td>14.1</td>
<td>31.4</td>
</tr>
<tr>
<td>DialogRE: (Yu et al., 2020)</td>
<td>1,578</td>
<td>20,724</td>
<td>13.1</td>
<td>249.6</td>
<td>19.1</td>
<td>49.2</td>
</tr>
<tr>
<td>FriendsQA (Yang and Choi, 2019)</td>
<td>969</td>
<td>16,415</td>
<td>16.9</td>
<td>307.1</td>
<td>18.2</td>
<td>50.1</td>
</tr>
<tr>
<td>DREAM (Sun et al., 2019)</td>
<td>345</td>
<td>1,706</td>
<td>4.9</td>
<td>74.3</td>
<td>15.2</td>
<td>32.1</td>
</tr>
</tbody>
</table>

| **Customer Service Dialogue**    |           |            |             |             |             |           |
| MetaLWOx (Lee et al., 2019)      | 37,860    | 431,580    | 11.4        | 146.0       | 12.8        | 45.2      |
| Schema (Rastogi et al., 2020)    | 22,559    | 453,320    | 20.1        | 322.7       | 16.1        | 65.1      |
| Taskmaster2 (Byrne et al., 2019) | 17,083    | 286,930    | 16.8        | 264.7       | 15.8        | 59.7      |
| MultiWOZ 2.2 (Zang et al., 2020) | 10,330    | 141,162    | 13.7        | 280.4       | 20.5        | 62.6      |
| MSR-E2E (Li et al., 2018)        | 8,769     | 68,406     | 7.8         | 150.0       | 19.2        | 46.9      |
| doc2dial (Feng et al., 2020)     | 4,662     | 62,581     | 13.4        | 268.1       | 20.0        | 53.8      |
| DREAM (Sun et al., 2019)         | 345       | 1,706      | 4.9         | 74.3        | 15.2        | 32.1      |

**Dialogue**

Peyton: I have been asking you to bring that video game for me.
Cameron: I am happy to deliver it to you.
Peyton: When would you come home?
Cameron: I have to stay out of town for another week.
Peyton: Can you just deliver that game through the courier?
Cameron: I don't have time to do that.
Peyton: Get the job done and come home then. ASAP.

ChatGPT (zero-shot)

Peyton asks Cameron to bring a video game, but Cameron is out of town and unable to come home. Peyton suggests delivering the game through courier, but Cameron refuses. Cameron will be out of town for another week and Peyton tells him to get the job done and come home ASAP.

BART-large

Cameron will have to stay out of town for another week. He will have to deliver the video game to Peyton.

MP4 (DAP-TOP)

Peyton wants Cameron to bring her a video game. Cameron will be out of town for another week.

**Ground Truth**

Peyton is expecting Cameron to bring the video game. Cameron will probably be out for another week.

**Dialogue**

Freddie: What are you watching on Netflix?
Kelly: I've just finished Mad Men.
Freddie: There should be new episodes coming soon.
Kelly: OMG! I can't wait! :(
Freddie: Try Outsider or The Tudors in the meantime.
Jim: I'm watching I, Zombie.
Bob: Try Grimm or Sabrina. They're both great!
Kelly: They are all horrible and disgusting!
Bob: I think they are funny. They're just TV shows, don't take them too seriously.
Greg: I'm watching House of Cards though I must admit the new series is boring.
Mike: Please don't. Not the same without Kevin Spacey! You should try The West Wing if you enjoy political dramas.
Greg: Netflix original Bodyguard is supposed to be good from what I've heard.
Mary: I'm a bit ashamed to confess I enjoy stuff like Pretty Little Liars or Gossip Girl.
Nancy: Don't be ashamed! I enjoy them too! And my favourite one is Desperate Housewives!!
Mary: I know! I watched it twice!
Alan: I'm catching up with Friends.
Nancy: OMG! I used to love Friends!

ChatGPT (zero-shot)

A group of people discuss their favorite shows on Netflix, recommend new shows to each other, and share their opinions on different genres. They mention shows like Mad Men, The Crown, Outlander, The Tudors, I, Zombie, Grimm, Sabrina, House of Cards, The West Wing, Bodyguard, Pretty Little Liars, Gossip Girl, Desperate Housewives, and Friends.

BART-large

Freddie has just finished watching Mad Men. Kelly has just finished The Crown. Jim is watching I, Zombie. Greg is watching House of Cards. Alan is catching up with Friends. Nancy's favourite series is Desperate Housewives.

MP4 (DAP-TOP)

Freddie, Kelly, Jim, Bob, Greg, Mary and Alan are discussing what they are watching on Netflix.

**Ground Truth**

Freddie, Kelly, Jim, Greg, Bob, Mike, Mary, Alan and Nancy are watching different shows on Netflix.